THE STATISTICAL THEORY OF BACTERIAL MUTATIONS

by J. B. S. Haldane, F.R.S.

Luria and Delbrück (1943) investigated the origin of bacteria resistant to phage. They found that after exposing about 10^9 *Escherichia coli* to phage the number of resistant bacteria remaining in culture was either zero, or a number small enough countable by plating out. They concluded, in my opinion entirely correctly, that the resistant had arisen by mutation. However it is possible that their statistical treatment could be improved, and it is the object of this paper to suggest possible improvements.

They started their cultures with a small number (about 50 to 500) of bacteria from a sensitive strain, added phage when the number had increased to between 10^8 and 4×10^{10} and then estimated the number of resistant bacteria by plating out a fraction of the culture, or in one case, the whole of it, and counting the colonies. They estimated the mutation rate either from the mean number of colonies, each representing one mutant, or from the frequency of cultures in which no survivors were found. The two methods gave different estimates , the former being usually greater by a factor of about 5, though both were of the order of 10^{-8}. It will be suggested that this divergence was due to their statistical method, though the theory here put forward lays no claim to finality, and is at best an advance in the right direction.

Theory of an Ideal Experiment

We shall first consider an ideal experiment, in which all individuals descended from a single sensitive bacterium, divisions are synchronous, and there are no deaths. Mutation is irreversible, and occurs at a division only one of the products being resistant.

Let: n be the number of generations.

$N = 2^n$ be the number of bacteria when phage is added.

m be the probability of mutation at a division.

x be the number of mutants in a culture of N.

p_x be the probability of finding just x mutants.

$g = 1/2\ mN$

Clearly N must be chosen so that g is of the order of unity. For if it is much smaller most cultures are sterile, if much larger they contain uncountable numbers of mutants. We shall first show how to calculate p for small values of x, and then how to calculate the moments of the distribution of x.

p_0 is the probability that no mutants are present. Then in none of $N-1$ divisions has a mutation occurred. Thus

$$p_0 = (1-m)^{N-1} = e^{-2g}\left(1+2g(1-g)N^{-1} + O(N^{-2}) \right)$$

So for moderate values of g, $p_0 = e^{-2g}$ with an error of less than 10^{-6}. Similarly p_1 is the probability that there has been one and only one mutation, and that this occurred during the last cycle of divisions, i.e., in one of $1/2\ N$ divisions out of a total of $N-1$ divisions. Thus

$$p_1 = m\ 1/2\ N\ (1-m)^{N-2}$$

2 $$p_1 = ge^{-2g}(1+2g(2-g)N^{-1} + O(N^{-2}))$$

To find p_x when x exceeds unity, we must consider all the partitions of x into powers of 2 including unity. The number of these partitions is the coefficient of t^x in the expansion of

$$\frac{1}{(1-t)(1-t^2)(1-t^4)(1-t^8)----}$$ in ascending powers of t. Each partition represents a set of mutations which could give rise to x mutants. Thus

$$5 = 2^2+1 = 2(2) + 1 = 2+3(1) = 5(1).$$

The pattern corresponding to each of these partitions is shown in Fig.1, mutant cells being represented by black, and normal by open circles.

2^2+1 $\quad$ $2(2)+1$ $\quad$ $2+3(1)$ $\quad$ $5(1)$

Fig. 1

Consider the mutations represented by the partition $2+3(1)$. One mutation occurred in one of the 1/4 N divisions of the penultimate set. The probability of such an event is m 1/4 N or 1/2 g. Three took place in the 1/2 N divisions of the last set. The probability is

$$1/3!\ m^3\ N/2\ (N/2-1)\ (N/2-2) \text{ or } 1/6\ g^3(1-6N^{-1}+O(N^{-2})).$$

The probability that 5 mutant bacteria originated in this particular way is thus $1/12\ g^4e^{-2g}(1-2(3-6g+g^2)N^{-1} + O(N^{-2}))$

Considering the other three possible partitions of 5 into powers of 2 we find:

$$p_5 = (1/4\ g^2 + 1/8\ g^3 + 1/12\ g^4 + 1/120\ g^5)\ e^{-2g} + O(N^{-1})$$
$$= \frac{g^5 + 10\ g^4 + 15\ g^3 + 30\ g^2}{5!\ e^{2g}} \qquad + O(N^{-1})$$

In general a partition of x into $a_k 2^k + a_{k-1} 2^{k-1} + \text{----} a_1 2 + a_0 1.$ gives rise to a term

$$\frac{g^{\Sigma a_k}\ e^{-2g}}{a_k!\ 2^{k a_k}\ a_{k-1}!\ 2^{(k-1)a_{k-1}}\ \text{-----}\ a_1! 2^{a_1} a_0!}$$

Table 1 gives the coefficients of $\frac{g^n}{x!}$ in the expansions of p_x up to x=10.

Thus $p_7 = \dfrac{630g^3 + 315\ g^4 + 105\ g^5 + 21\ g^6 + g^7}{7!\ e^{2g}}$

Table 2 gives values of p_x for several values of g. It will be seen that while there is some analogy with the Poisson distribution there is no single maximum value when g is of the order of unity. These values are only strictly accurate when N is infinite. But the are reasonably accurate when x is small compared with $N^{1/2}$. Thus the last term of the expansion of p_{100} is

$$\frac{g^{100}}{100!\ e^{2g}}\ (1-(g^2-100g+4950)N^{-1} + O(N^{-2})).$$

Unless g exceeds 200, the error is less than 2.5×10^{-4} when $N=10^8$. The moments cannot be calculated by this method, for as N tends to infinity, all the moments of x do so. For when $x=2^k$ the leading term in the expansion of p_x is $2^{-k} ge^{-2g}$, whilst otherwise the coefficient of ge^{-2g} is zero. Hence when N tends to infinity the coefficient of ge^{-2g} in $\bar{x}$ or $\Sigma x p_x$ tends to 1+1+1+1---, which diverges. Hence $\bar{x}$ and the higher moments diverge. The moments must be calculated by a different method when N is finite.

When $N=2^n$, let $\bar{x}=u_n$, $\bar{x^2}=v_n$, $\bar{x^3}=w_n$. Now consider what happens if one more generation is added at the beginning, so that $N=2^{n+1}$.
In 1-m of all cases there is no mutation at the first division. Hence we simply have two populations of 2^n bacteria each derived from a susceptible individual. If between them they contain x' mutants, then the cumulants of the distribution of x' are twice those of the distribution of x, and the moments are derivable from them by well-known expressions; so

$$\bar{x}'=2\bar{x},\ \overline{x'^2}=2\bar{x^2}+2\bar{x},\ \overline{x'^3}=2\bar{x^3}+6\bar{x}.\bar{x}^2,\ \text{etc.}$$

But in a fraction m of all cases a mutation occurs in the first division, and $x'=2^n+x'$, whence

$$\bar{x}'=2^n+\bar{x},\ \overline{x'^2}=2^{2n}+2^{n+1}\bar{x}+\bar{x^2},\ \overline{x'^3}=2^{3n}+3\cdot2^{2n}\bar{x}+3\cdot2^{n}\bar{x^2}+\bar{x^3},\text{etc.}$$

Hence
$$u_{n+1}=(1-m)2u_n+m(2^n+u_n)$$
$$=(2-m)u_n+2^n m$$

(3) So $2^{-n-1}u_{n+1}-2=(1-m/2)\,(2^{-n}u_n-2)$, whence

$$2^{-n}u_n-2=(1-m/2)^n(2^{-n}u_0-2)$$
$$=-2(1-m/2)^n,\ \text{or}$$
$$u_n=2(2^n-(2-m)^n)$$
$$=2^n-1/4\ 2^n m^2 n(n-1)+\text{------}$$
$$=2ng-1/2\ mn(n-1)g+O(m^2).$$

Thus $\bar{x}=u_n=2g\log_2 N-g^2\log_2 N(\log_2 N-1)N^{-1}+O(N^{-2})$.

The second and later terms are negligible.

Similarly
$$v_{n+1}=2(1-m)(v_n+u_n)+m(2^{2n}+2^{n+1}u_n+v_n)$$
$$=(2-m)v_n+2^{2n}m+2(2^n m+1)2^n mn+O(m).$$

Hence
$$\frac{v_{n+1}}{(2-m)^{n+1}}-\frac{v_n}{(2-m)^n}=\frac{2^{2n}m+2(2^n m+1)2^n mn}{(2-m)^{n+1}}+O(m^2)$$

or $$\frac{v_n}{(2-m)^n} - \frac{v_{n-1}}{(2-m)^{n-1}} = m(2^{n-2}+n-1) + 2^{n-3}m^2(5n-9) + O(m^2).$$

Summing this expression

$$\frac{v_n}{(2-m)^n} = m(2^{n-1}+1/2\ n(n-1)) + m^2(5n-9)2^{n-2} + O(m^2)$$

or $$v_n = 2^{n-1}m(2^n+n(n-1)-2) + 2^{2n-2}m^2(6n-9) + O(m)$$

i.e. $\bar{\bar{x}}^2 = v_n = Ng + (n^2-n-2)g + 3(2n-3)g^2 + O(N^{-1})$

Similarly $w_{n+1} = 2(1-m)(w_n+3u_nv_n) + m(2^{3n} + 3\cdot 2^{2n}u_n + 3\cdot 2^{2n}v_n+w_n)$

whence $\bar{X}^3 = 1/2\ N^2g + O(N)$.

X^3 is of order N^{n-1}, and the momen ts of x about its meanare of the same order as those about zero, as are the cumulants of its distribution.

To sum up, the mean, variance and third momen t of x are:

$$k_1 = \frac{2g \log N}{\log 2}$$

$$k_2 = gN$$

$$k_3 = 1/2\ gN^2$$

Within an accuracy of the order of N^{-1}.

This implies that the mean cannot be used with any confidence for the estimation of m. For suppose we have values of x from S populations of N, the variance of the mean is S^{-1} that of x, and the third moment S^{-2} that of x. For example if $N = 2^{2n} = 1.342\times10^8$, and $S=100$, while $m = 1.5\times10^{-8}$ or $g=1$, then the mean of x is 54, and the variance of the mean is 1.342×10^6, so that its "standard error" is 1159, and the measure of skewness γ_1 is about $1/2\ (N/g)^{1/2}$, or 12,000. Thus the mean will almost always be less than its expected value, and no practicable number of experiments will enable us to estimate m from it.

On the other hand in this case $p_0=p_1=1354$ and a satisfactory estimate of m can be obtained from the number of cultures containing one mutant or none. It is worth noting that here, probably for the first time in the history of science, we have to deal with a distribution similar to that of the gains of a gambler in Euler's famous Petersburg problem, in which the gambler produces a "head" n times in succession. The moments are finite if the banker has a finite but large capital of 2nN roubles.

Samples From an Ideal Experiment

Luria and Delbrück usually worked with a fraction of their total population, which they plated out. Suppose a fraction c is taken, and P_y is the probability that it contains y mutant bacteria . Then

$$P_0 = p_0 + (1-c)p_1 + (1-c)^2 p_2 + \text{------} (1-c)^x p_x + \text{----------}$$

$$P_1 = c(p_1 + 2(1-c)p_2 + \text{---------} x (1-c)^{x-1} p_x + \text{--------------})$$

$$P_2 = c^2(p_2 + 3(1-c) p_3 + \text{----------} 1/2 \cdot x(x-1)(1-c)^{x-2} p_x + \text{------})$$

$$\text{---}$$

$$P_y = c^y (p_y + (y+1)(1-c)p_{y+1} + \text{--------} \binom{y+n}{n}(1-c)^n p_{y+n} + \text{----------})$$

It follows that

$$p_0 = P_0 - \left(\frac{1-c}{c}\right) P_1 + \left(\frac{1-c}{c}\right)^2 P_2 - \text{------} + \left(\frac{c-1}{c}\right) P_x + \text{------------})$$

$$p_1 = c^{-1}(P_1 - \left(\frac{1-c}{c}\right) 2P_2 + \left(\frac{1-c}{c}\right)^2 3P_3 + \text{--------} + \left(\frac{c-1}{c}\right) P_x + \text{-------})$$

$$\text{---}$$

$$p_r = c^{-n}(P_r - (r+1)\left(\frac{1-c}{c}\right) P_{r+1} + \text{------} + \binom{x}{r}\left(\frac{c-1}{c}\right)^{x-n} P_x + \text{-------})$$

Hence no simple expression can be given for P_y as a function of y, even when y=0.

The fractional moments of y are simple fractions of those of x , namely

$\bar{y} = c\bar{x}$, $\overline{y(y+1)} = c^2\overline{x(x-1)}$, etc.

Hence $\bar{y} = 2cgn$

$$\overline{y^2} = c^2 gN + O(1)$$

$$\overline{y^3} = 1/2\, c^3 gN^2 + O(N), \text{ etc.}$$

For the reasons given above these moments are almost useless for estimating m. Unfortunately where $c = .05$, as in many of Luria and Delbrück's experiments, the series for P_y converge so slowly that they are of little value.

Deviations from the Ideal

An experiment may deviate from the ideal form for a number of reasons of which we shall consider three. In the first place, divisions are not quite simultaneous throughout the culture. With an average of say 27 divisions, many bacteria will be the product of 26 or 28 some perhaps of 17 or 37 divisions. However, if all bacteria survive, there have been exactly N-1 in all. Hence the value of p_0 is unchanged. Further, the probability that a mutation occurred during the origin of any of the N bacteria is exactly 1/2 m. So the value of p_1 is unchanged.

On the other hand the value of p_2 is diminished. For the terminal sections of the pedigree are all of one or other of the types shown in Fig. 2

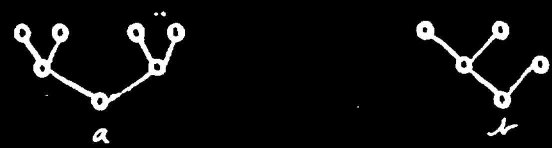

Fig. 2

Let a and b be the numbers of these two types. Then $N = 4a + 3b$. Now two mutants may occur in situ because either product of the first division of the first type is an original mutant, but only because one of the two products of the first divisin of the second type is an original mutant.

A pair of non-sib mutants can occur just as before. Hence

$$p_2 = (\,m(a+1/2\ b) + m^2/2\ (N/2)^2)\ (1-m)^N + O(N^{-1})$$
$$= \left(\left(\frac{2a+b}{4a+3b}\right)g + 1/2\ g^2\right) e^{-2g}, \text{ approximately}$$

Hence $(1/2\ g + 1/2\ g^2)e^{-2g} \geqslant p_2 \geqslant (1/3\ g + 1/2\ g^2)e^{-2g}$

Thus the first term in the expansion of p_2 is in general diminished, though not very greatly. On the other hand the first term in that of p_3 is no longer zero. Hence the p_x series is smoothed out and its terms ~~have~~ no longer have such sharp maxima when x is a power of 2 and g is small. In consequence it is only legitimate to use p_0 and p_1 in the estimate of g and m.

A second case of deviation is the case of death of some bacteria. The probability that a given bacterium will die before it divides will vary with time, but we may assume it to have a constant value h, as it is only the value during the last few generations that matters. The mean number N of bacteria produced by n synchronous divisions is $2^n(1-h)^n$. Thus $n = \dfrac{\log N}{\log 2 + \log(1-h)}$. The number of last divisions is $N/2(1-h)$ if we assume that death occurs just after a division, and 1/2 N if we assume that death occurs just before. If deaths are evenly spread out it is about $N/2-h$. The number of penultimate divisions is $N/2(1-h)(2-h)$, of antepenultimate divisions $N/4(1-h)^2(2-h)$, and so on, the total being $2(1-h)N/(2-h)(1-2h)$. This excludes divisions all of whose progeny die out. Hence

$$p_0 = \left(1 + gh/1-h + g(1+g)h^2/2(1-h)^2 + \text{-----}\right) \operatorname{Exp} -(-4(1-h)g/(2-h)(1-2h)$$
$$= \left[e^{-2g(1+h+O(h^2))}\right].$$

and similar expressions may be obtained for p_1 etc. Hence the effects of deaths will be to lessen p_0 for given values of N and m, and hence to lessen the value of m for given observed values. It is impossible to allow for their effect exactly unless it is known how they are distributed through the bacterial life cycle.

Thirdly, it is clear, since phage resistant mutants are rare, either that resistant types are under some disadvantage compared with normal, or that the rate at which they mutate back to susceptibility is much greater than m. The latter alternative has a negligible effect on p_x when N is small, though it makes moments finite however large is the value of N. Suppose the time taken before a mutant divides is q^{-1} times the time taken by a normal, i.e. the mean growth rate of mutants is q times that of normals. Consider what happens when in the absence of mutation there would be 2^n bacteria. If x=o on 1, this number will be unaltered and p_o will be unchanged. But p_1 is increased because only a fraction q of mutants in what would otherwise have been the penultimate generation will have divided. So

$$p_1 = (1 + 1/2\, q)\, g\, e^{-2g},$$

and there will be similar changes in other values of p_x. If q is measurable it may be possible to make the necessary allowance. It can also be shown that the moments are finite. For example the recurrence equation for $u_n = \bar{x}$ becomes $u_{n+1} = 2(1-m)u_n + m(u_n + 2^n q^n)$, whence $\bar{x}$ approximates to $g/1-q$ when n is large. To sum up this section, the only influence likely to upset p_o is a heavy mortality. p_1 can also be upset by differential growth rate but not by non-synchronous division. This latter cause will upset other values of p_x.

The Estimation of the Mutation Rate

It follows from the above discussion that the mean number of mutants in a series of cultures will give a very unreliable estimate of m. The estimate least liable to be affected by the various causes of deviation is made FROM the frequency of cultures containing o and 1.

If we are going to rely only on the number of sterile cultures, suppose S cultures are examined and a prove sterile, then it is easy to show tha the best estimate of g is

$$\hat{g} = 1/2 \log_e (s/a) \pm 1/2 ((s-a)/sa)^{1/2} \quad \text{--------(1)}$$

If there are a cultures with no mutants and b with one, the logarithm of the likelihood is

$$L = b \log g + (s-a-b) \log(e^{2g} - g - 1) - 2sg$$

$$\frac{dL}{dg} = \frac{b}{g} + \frac{(2e^{2g}-1)(s-a-b)}{e^{2g}-g-1} - 2s, \text{ whence}$$

$$e^{2g} = \frac{2sg^2 + (s+a)g - b}{2(a+b)g - b} \quad \text{--------------------(2)}$$

gives the maximum likelihood estimate of g. The sampling variance of this estimate is

$$\left(\frac{ge^{2g}(e^{2g}-g-1)}{(4g^2+1)e^{2g}-1} \right)$$

or approximately

$$\frac{b(s-a-b)}{s(a^2+4b^2)-a^3}$$

A Numerical Example

In their experiment number 23, Luria and Delbrück tested 87 cultures containing $N=2.4\times10^8$ bacteria apiece. After adding phage, 29 had no survivors, 17 had one, 4 had two, etc. (See my Table 3) Four cultures had 201-500 survivors, and the mean was 28.6.

. S=87, a=29, b=17.

So equation (1) gives g=.549 ± .076
and equation (2) gives g=.542 ± .068
whence $m=4.8\times10^{-9}$

Luria and Delbrück derive 4.7×10^{-9} from Poisson theory, and

24.5×10^{-4} from the mean. For the reasons given above, the latter estimate is of little value. Since $n=\log_2 N = 27.83$, the expected value of the mean is 30.16, which is very close to the value found, but the agreement is fortuitous.

The number of cultures a_x in which x mutants were found are given in Table 3. The fit of a_2 and the higher valuesis very bad. $X^2 = 19.7$ for 5 degrees of freedom. On the other hand if we only consider a_0 and a_1 (i.e. a and b), $X^2 = .09$ for one degree, a very good fit. Thus the results are compatible with the view that the only important deviations from ideality are due to non-synchronous divisions.

In their other experiments these authors only counted a portion of the culture. Hence m' could only be estimated if thedistribution held exactly for high values of x. It can best be estimated on this hypothesis when P_0 is fairly large. In Experiment 21a they plated out 1/4 of their culture of $N= 1.1\times10^8$ bacteria, and found $P_0 = 11/19$ or .579. The estimate of g is about .55, giving $m = 10^{-8}$, approximately.

In experiment 16, c=.4, $P_0=11/20$, $N=5.6\times10^8$, in expt. 17, c=.4, $P_0= 5/12$, $N = 5\times10^8$, hence $m= 1.8\times10^4$ in the first case, and about 1.5×10^9 in the second. These values are uncertain, owing to the large influence of p_2 and higher terms.

Table I

x \ n	1	2	3	4	5	6	7	8	9	10
1	1									
2	1	1								
3	0	3	1							
4	6	3	6	1						
5	0	30	15	10	1					
6	0	90	1105	45	15	1				
7	0	0	630	315	105	21	1			
8	5040	1260	1260	2625	840	210	28	1		
9	0	45360	11340	11340	8505	2016	348	36	1	
10	0	226800	283500	75600	57645	23625	4410	630	45	1

Coefficients of $\frac{g^n}{x!}$ in the expansion of p_x

Table II

g \ x	0	1	2	3	4	5	6	7	8	9	10	10+
.1	.8187	.0819	.0430	.0042	.0217	.0022	.0011	.0001	.0105	.0011	.0006	.0229
.2	.6650	.1330	.0798	.0142	.0379	.0074	.0042	.0007	.0177	.0035	.0010	.0355
.5	.3679	.1839	.1380	.0536	.0525	.0327	.0199	.0072	.0290	.0140	.0100	.0913
1	.1354	.1354	.1354	.0902	.0902	.0632	.0481	.0288	.0418	.0296	.0253	.1768
2	.0183	.0366	.0544	.0611	.0672	.0659	.0627	.0566	.0397 .0506	.0396 .0392	.0346	.4488

p_x in terms of g and x

Table III

r	0	1	2	3	4	5	5+
a_r obs	29	17	4	3	3	2	29
a_r calc.	29.43		12.30	5.10	5.08	2.97	
		15.95					15.6

Luria and Delbrück's experiment 23